DATA SHARING WORKBOOK

INTRODUCTION

Scientists working in many different areas are already sharing their data through a variety of mechanisms. However, some disciplines are less familiar with this process and associated practices. **The goal of this workbook is to show how investigators working in a variety of scientific areas have shared their data.** To highlight the benefits of data sharing, we have included testimonials from investigators who are already sharing their data.

NIH supports a wide range of scientific research. Some studies, such as small laboratory-based projects, make raw data available in publications. These studies generally are based on small numbers of laboratory animals, specimens, or clinical subjects. Publishing the raw data constitutes an acceptable mechanism for sharing data, provided that privacy of human subjects is protected. However, raw data from large studies are not amenable to sharing through publication. Such studies can make data available through data archives or enclaves. For example, X-ray crystallography, gene mapping, and survey data are available from data archives or repositories, some with sophisticated Web interfaces. Data from human subjects present special concerns regarding data sharing. The rights and privacy of individuals who participate in NIH-sponsored research must be protected at all times, and patentable and other proprietary data should also be protected.

In summary, all data should be considered for sharing. Data that constitute "unique resources" especially should be shared unless there is a strong reason not to. Such data are difficult if not impossible to replicate because of cost (e.g., large national longitudinal surveys), special circumstances (e.g., health effects associated with a natural disaster), or rare population (e.g., a sample of centenarians). Less likely candidates for sharing are data from small studies involving research procedures that are easily replicated or data from human subjects that might identify them.

PROTECTING THE RIGHTS AND PRIVACY OF HUMAN SUBJECTS

An important issue associated with the sharing of all data derived from human subjects is the protection of research participants' identities. The rights and privacy of people

who participate in NIH-sponsored research must be protected at all times. Sensitive data raise special concerns about confidentiality and the protection of subjects' privacy because of a greater possibility of harmful social, economic, or legal consequences if released. However, the collection of sensitive data does not preclude sharing. Indeed, some of the examples of sharing highlighted below include items on highly sensitive and, sometimes, illegal behaviors. But sensitive data call for a higher level of security during collection, analysis, and storage and special consideration when preparing datasets for broader use.

What constitutes "sensitive" data varies by context, population, and time. Illegal and sexual behaviors are almost always considered sensitive. Measures of alcohol use are less sensitive among adults than underage adolescents. Many diseases and medical conditions, such as bipolar illness or HIV infection, could be considered sensitive information. Because access to health insurance and employment can be affected by pre-existing conditions or even risk for certain diseases, information about medical conditions as well as genetic markers and family history, which may be used as indicators of predisposition, are also considered to be sensitive information.

There are two basic tools to protect from disclosure of sensitive data and subjects' identities: Restricting information in the dataset, and restricting access to the data. Thus, data intended for broader use should be free of identifiers that would permit linkages to the research participants and free of content that would create unacceptably high risks of subject identification.

Stripping a dataset of items that could identify individual participants is referred to by several different terms, such as data redaction, de-identification of data[1], and anonymizing data. It is rarely sufficient to simply remove names, addresses, telephone numbers, Social Security Numbers, and the like. Deductive disclosure of individual subjects becomes more likely when there are unusual characteristics or the joint occurrence of several unusual variables. Samples drawn from small geographic areas, rare populations, and linked datasets can present particular challenges to the protection of subjects' identities.

[1] Under the HIPAA Privacy Rule, de-identification of a dataset means removing the following variables: names; geographic information (including city, state, and zip code); elements of dates such as those for birth, hospital admission and discharge, death; telephone numbers; fax numbers; electronic mail addresses; Social Security Number; medical record and prescription numbers; health plan beneficiary number; account numbers; certificate or license number; any vehicle identifier or serial number, including license plate number; any device identifier or serial number; Web Universal Resource Locator (URL); Internet Protocol (IP) address number; any biometric identifiers, including finger or voice prints; full face photographic images or any comparable images; and any other unique identifying number, characteristic, or code consisting of any segments of the previously listed identifiers.

There are many other methods currently used to anonymize data. Some investigators withhold parts of the sample; others block access to specific variables, especially items with low prevalence rates that make it easier to identify participants with unusual characteristics. Scientific interest in protecting subject identity is growing, and new methods are actively being developed. For example, investigators are creating synthetic datasets that mimic the characteristics of the original dataset without risking the identification of individual participants. It is beyond the scope of this document to instruct investigators about methods used to protect the identity of research subjects. However, several references are provided below. Investigators should also consult with statisticians to determine the best plan for data redaction and test the redaction process prior to the release of data.

Measures used to minimize the risk of breaching the confidentiality of data include the following:

- Mandatory agreements to maintain confidentiality
- Data encryption
- Electronic firewalls and locked storage facilities,
- Password authentication of users
- Audit trails
- Disaster prevention and recovery plans
- Security measures for backup tapes.

Institutions and investigators should work closely to develop and update plans and procedures to protect the security of data.

Data-use sharing agreements put limitations on who can use the data and how they are to be used. (These documents are also known by other names, such as license agreements, data-distribution agreements, and data-sharing agreements.) Such agreements contain different types of requirements, including those to protect the privacy of subjects and the confidentiality of the data. These documents can incorporate confidentiality standards to ensure data security at the recipient site and prohibit manipulation of data for the purposes of identifying subjects. They can stipulate that the recipient not transfer the data to other users, that the data are only to be used for research purposes, that the proposed research using the data will be reviewed by an IRB, and the like. Penalties for violating terms of the agreement are generally specified in these agreements. Below we describe some of the terms included in data-use sharing agreements used by archives and other entities that have shared data.

PROTECTING PROPRIETARY DATA

NIH encourages sharing of data generated with its support for further research, development, and application in the expectation that this will lead to products and knowledge of benefit to the public. However, NIH recognizes the need to protect patentable and other proprietary data and the restrictions on the sharing of data that may be imposed by agreements with third parties. In this regard, note that under the Bayh-Dole Act, grantees have the right to elect and retain title to subject inventions developed with Federal funding. Indeed, for inventions developed in its intramural

program, NIH does file patent applications in accord with a set of policies that are described at http://www.nih.gov/od/ott/200po6.htm. It is not the intent of the NIH statement on data sharing to discourage, impede, or prohibit the development of commercial products from federally funded research. However, it should be noted that, in general, NIH does not support the production of data that cannot be shared. If patent protection is being sought, data still can be shared in a timely manner.

EXAMPLES OF DATA SHARING

Data Archives

There are many archives for data. Many data archives facilitate the sharing of data using Web-based platforms. A searchable list of Websites for archives is available through the University of California at San Diego at http://odwin.ucsd.edu/idata/.

Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication. The **National Center for Biotechnology Information (NCBI)**, National Library of Medicine (NLM), NIH, was established in 1988 as a national resource for molecular biology information. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information with the goal of improving understanding of molecular processes affecting human health and disease. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit. For more information about submitting and downloading data, see the NCBI Website at http://www.ncbi.nlm.nih.gov/Genbank/index.html

The National Center for Chronic Disease Prevention and Health Promotion at CDC operates the **Youth Risk Behavior Surveillance System (YRBSS)**. This system provides data on six health risk behaviors among youth: unintentional injuries and violence, tobacco use, alcohol and other drug use, sexual behaviors, dietary behaviors, and physical activity. The YRBSS is composed of several surveys of different populations of youth, but focuses on national, State, and local school-based surveys of students in grades 9 through 12.

National YRBSS data are available directly through the Internet at the CDC Website. See http://www.cdc.gov/nccdphp/dash/yrbs/. Data from each wave of the national survey can be downloaded along with documentation files at http://www.cdc.gov/nccdphp/dash/yrbs/data/index.htm The data files for the most recent wave are available in ASCII, SAS, and SPSS. Documentation files are in PDF. User service is available by telephone and email. The YRBSS Website also contains a copy of the current questionnaire, item rationale, and results from previous waves. In addition, users can request a free CD-ROM with 6 years of compiled YRBSS data from the national, State, territorial, and local school-based surveys.

The CDC minimizes the risk of inadvertent disclosure of subjects by collecting the data anonymously. Participants complete a self-administered questionnaire in their regular classroom settings. Only four demographic variables are measured: Age, grade, race/ethnicity, and gender. School and classroom codes are not included in the datasets, so it is not possible to determine the school in which a student was enrolled.

Another example is **Sociometrics Corporation** (http://www.socio.com/), which maintains over 300 datasets from 200 different studies in seven topical areas: AIDS and other sexually transmitted diseases, disability, the American family, aging, adolescent pregnancy and pregnancy prevention, maternal drug abuse, and contextual influences on behavior. This archive has been in operation for more than 15 years. An expert panel selects the datasets included in Sociometric's library on the basis of scientific merit, substantive utility, technical quality, and potential for secondary analyses. For the cost of the data (approximately $100 to $225 per dataset if purchased individually, less if the entire collection is purchased), Sociometrics provides a complete data file in both CD-ROM and Internet formats, SPSS and SAS program statements, search and retrieval software, data summaries, detailed users' guides, and technical assistance.

The purchaser of data from Sociometrics can be either an individual or an institution. If the purchaser is an institution, an institutional representative must sign a license agreement certifying that only faculty, students, and staff can use the data. The license agreement further stipulates that neither printed nor electronic data may be copied or otherwise shared. Use of the data is restricted to statistical reporting, analysis, and teaching. The agreement prohibits the user from making any efforts to identify individual cases and prohibits linking data from this archive with individually identifiable data from other datasets. Violation of the license agreement carries civil liability.

The **Inter-University Consortium for Political and Social Research at the University of Michigan** has prepared an excellent set of guidelines for preparing data for archiving. While these guidelines were written with social science data in mind, they are broadly applicable. For further information, see http://www.icpsr.umich.edu/

Data Enclaves

Some data can be shared only under the most controlled conditions. If, for example, there is any risk of subject identification, the investigator may ask that users submit requests for specific analyses or come to the investigator's site to run analyses under supervision. Data enclaves were designed to deal with such situations.

One such enclave is the **Research Data Center at the CDC's National Center for Health Statistics (NCHS)**. The Research Data Center supports use of several NCHS restricted-use datasets through the Internet and within the Data Center itself. Additional information on the Research Data Center is available at http://www.cdc.gov/nchs/r&d/rdc.htm.

One of the datasets that can be used at the NCHS Research Data Center is a periodic survey called the National Survey of Family Growth (NSFG). Data from this survey provide an accurate statistical picture of family life, marriage and divorce, contraception, sexual experience, pregnancy, and infertility. Information concerning the NSFG is available at the NCSH Website at http://www.cdc.gov/nchs/nsfg.htm.

NCHS encourages the use of data from the NSFG, which are available through a variety of mechanisms. Public use datasets of the NSFG and other NCHS datasets are available free or at minimal cost after signing a use agreement. However, some NSFG data are not released in order to protect participants' identities. The restricted data, which are referred to as the NSFG contextual data file, do not include direct identifiers, such as name or social security numbers, but they may contain codes for small geographic units, such as blocks or census tracts. Thus, the restricted contextual dataset is only available to approved researchers via a remote access procedure or for analysis at the NCHS facility in Hyattsville, Maryland.

In order to gain access to restricted data, researchers must submit a detailed description of their projects. The proposal must include personal identification and institutional affiliation, a current resume, dates of proposed tenure at the data enclave, source of funding, a detailed summary of the proposed research including a statement of why publicly available data are insufficient, and a complete list of data requested, including data system, files, years, variables, and the like. NCHS staff are available for consultation on the proposal development. A committee consisting of NCHS staff, including the Confidentiality Officer, reviews all proposals. This review addresses the following critical questions: Does the proposed activity constitute statistical research or an illegal attempt to identify respondents? If it is research, is there any risk that respondents will be identified inadvertently?

All applicants are also required to sign an agreement of confidentiality. This agreement prohibits copying files or portions of files, keeping restricted materials, attempting to learn the identity of participants, removing any printouts, electronic files, or other documents from the enclave unless authorized by NCHS staff. In addition all papers or reports submitted for publication must first be submitted to NCHS for disclosure limitation review.

The fee charged for work at the data enclave ($200 per day or $1,000 per week) includes space, equipment, staff time for supervision and disclosure limitation review, and the creation and maintenance of data files required by the researcher. All work must be completed within the confines of the enclave. No electronic or hard copies of data can leave the facility unless they are submitted to a disclosure limitation review. In addition, researchers must work under the supervision of NCHS staff during normal working hours.

It should be noted that data collected by NCHS are protected by the Public Health Service Act (Section 308(d)). Under this section, identifying data can be disclosed or used for a purpose other than that for which it was supplied only if the person or establishment identified has consented.

Mixed Mode Sharing

The **National Institute of Mental Health (NIMH) Human Genetics Initiative** collects and distributes family data on schizophrenia, bipolar disorder, Alzheimer's disease, and other mental disorders. Through the Initiative, qualified investigators can request clinical data, DNA samples, cell line cultures, and data derived from genotyping and other genetic analyses. Information on this initiative is available at http://zork.wustl.edu/nimh/NIMH_initiative/NIMH_initiative_link.html

Researchers can gain access to these data by successfully competing for an NIMH award specifically to analyze these data or by submitting an access request if they have no such award. Access certification is made on the basis of the experience and the scientific qualifications of the investigator. Requests must be submitted in writing on the letterhead of the sponsoring institution at which the research will be conducted and should include identifying information about the Principal Investigator and Coinvestigators, including curricula vitae. If access certification is obtained, biomaterials (DNA or cell lines) can be obtained at cost from the NIMH Center for Genetic Studies (http://zork.wustl.edu/nimh/). The Center serves as a data repository and management facility maintained under an NIMH contract.

All investigators must complete a Distribution Agreement, which includes a description of the research project to be conducted. The PI and an authorized representative of the recipient institution must sign the Distribution Agreement. The Agreement specifies that the investigator will only use the data and biomaterials for the specific project as described. The Agreement is not transferable to another recipient or facility, and biomaterials and data may only be shared with others by obtaining them directly through the NIMH center. The recipient must also agree to not attempt to establish the individual identities of subjects who provided the data or biomaterials.

When an access certification has been approved or a grant awarded, pedigree drawings are sent to the PI. Electronic files of clinical and genetic data and other information are available through password protected Websites. The investigator specifies which biomaterials are desired and sends this list to NIMH, which forwards it to the NIMH Center for Genetic Studies. The Center then provides shipping and payment instructions. All biomaterials and clinical data are stripped of personal identifiers. (No personal identifiers are ever received or handled by the Center.) The Center also provides periodic updates of data. Recipients share with the Center all genetic analysis data that they generate within 12 months of receipt of biomaterials or upon publication of research findings, whichever comes first. Upon completion of the project, the recipient must return all biomaterials as well as clinical and genetic data received from

the Center or certify that the clinical and genetic data were destroyed in accordance with applicable laws and safety procedures.

The National Longitudinal Study of Adolescent Health (Add Health) is a very large national survey of students in grades 7 through 12. Data for this longitudinal study were collected in three waves. The first wave included questionnaires completed by 90,000 students and in-depth interviews with 20,745 adolescents. The resulting dataset includes items on a range of characteristics and behaviors of adolescents, including sensitive behaviors, such as alcohol use and sex.

From the outset of this study, the investigators planned to share the data. Critical to the protection of subjects is the separation of identities from the data, which occurs immediately after data collection. Only a Security Manager can link the name and address of respondent to interview data. The investigators also asked for and received a Certificate of Confidentiality from DHHS to protect subjects' identities. All Add Health staff are required to take training in data confidentiality and security issues. Individuals and institutions seeking to obtain the Add Health data are encouraged to develop and implement a similar training program. Details about accessing the data are available on the study's Website at http://www.cpc.unc.edu/addhealth/. Only certified researchers are permitted access to Add Health data. Thus, the informed consent document notes that the study "is helping researchers understand the health of young adults and the behaviors that affect their health."

Because of the extremely sensitive nature of some of the information collected, the investigators have made data or portions of the data available in three ways: (1) a public use dataset that can be accessed through a data archive; (2) a restricted access contractual dataset, and (3) access at a data enclave at the Add Health facility under the supervision of staff.

The public use dataset includes only a subset of respondents to protect the identity of participants. The investigators found that even with more than 90,000 cases, a cross-tabulation of 5 variables could distinguish an individual record. Therefore, half of the sample was chosen at random for the public use dataset with an oversample of minority adolescents. CD-ROMs are distributed through a data archive run by Sociometrics Corporation (http://www.sociometrics.com/). The data are in ASCII format and can be analyzed with several standard statistical packages.

The restricted access dataset is available only to certified researchers who provide a nonrefundable fee to cover administrative handling charges and user support. Add Health investigators have embedded a hidden signature identifying the purchaser in each electronic file, so that unauthorized copies can be traced. All users must sign an agreement to maintain privacy of subjects and confidentiality of the data. In addition, users must certify that they have complied with a set of security requirements covering how the data are handled and stored. These requirements are updated periodically to reflect changes in computer technology. Applicants are also required to submit letters from their IRBs verifying and approving plans for data security and for minimizing risks

of deductive disclosure. The staff from Add Health conducts site visits to monitor the use of these data at outside institutions. The user fee covers the cost of these visits.

Researchers requesting use of data that cannot be shared through contractual agreements must come to the Add Health site at the University of North Carolina in Chapel Hill to conduct analyses under the supervision of Add Health staff. Again, these data are only available to certified researchers.

TESTIMONIALS

Unrestricted access to the mouse sequence should enhance efforts to identify causative genes in mouse models of diseases as well as identify human genes responsible for various disorders. The rapid progress toward making these data widely available will in turn speed the research for new ways to treat or even prevent disease.
Arthur Holden, Chairman of the Mouse Sequencing Consortium

We placed our NSAM data in the Data Archive on Adolescent Pregnancy and Pregnancy Prevention where it has been accessed for analysis and classroom projects. The advantage of this approach is that the Archive provides technical assistance and publicity on data availability to the research and education communities.
Freya Sonenstein, The Urban Institute
PI, National Survey of Adolescent Males (NSAM): 1988, 1990-91, 1995

The data were placed in the public domain, free of charge, six months after completing the last interview. Within a year, researchers at over 100 institutions were using the data." Over 300 published articles have used these data.
Larry Bumpass and James Sweet, Center for Demography and Ecology, University of Wisconsin-Madison
National Survey of Family and Households: 1987-99, 1992-94

We place our cleaned files on the web for free dissemination. There is a vast group of users, particularly for our China and Russia surveys. The culture of empirical research is new to Russia, and our surveys have helped to revolutionize social science training in that country. Virtually every Russian MS or PhD in microeconomics uses our Russian surveys for their work and these data are used by most scholars in that country along with about 800 other institutions around the world.
Barry Popkin, Department of Nutrition and Carolina Population Center, University of Carolina at Chapel Hill
Cebu Longitudinal Health and Nutrition Surveys: 1982-1986
China Health and Nutrition Survey: 1989, 1991, 1993, 1997, 2000
Russian Longitudinal Monitoring Surveys: 1991-2000

I have shared data from my research on motivation in the pigeon. Overall, I found the benefits of sharing to greatly outweigh the bother of pulling together the data and annotating them. I have also benefited from receiving data from colleagues along with programs vital to appropriate analyses. I truly believe that data sharing is an important component of the research enterprise.
Peter Killeen, Department of Psychology, Arizona State University